FMI 2014 - Free Models Initiative

Workshop Proceedings

Störrle, Harald; Hebig, Regina; Knapp, Alexander

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

GI Modellierung 2014

March 19 – 21, 2014 • Vienna (Austria)

FMI 2014 - Free Models Initiative

Workshop Proceedings

Harald Störrle, Regina Hebig, Alexander Knapp (Eds.)

Published on Sept 2014 v1.0

DTU Compute Technical Report -2014-14
Matematiktorvet
DK 2800 Kongens Lyngby

Editors' addresses:

Harald Störrle,
Technical University of Denmark (Denmark)

Regina Hebig,
Sorbonne Universités, UPMC Univ. Paris 06, LIP6, Paris (France)

Alexander Kanpp
University of Augsburg (Germany)

Table of Contents

The Free Models Initiative 2
Harald Störrle, Regina Hebig, Alexander Knapp

Experiences on the Quality and Availability of Test Models for Model Differencing Tools 11
Pit Pietsch, Dennis Reuling, Udo Kelter, Jens Folmer, Birgit Vogel-Heuser

The Need for Process Model Corpora 14
Tom Thaler, Jürgen Walter, Peyman Ardalani, Peter Fettke, Peter Loos

Towards an Open Process Model Repository for Evaluations in Business Process Management Research 18
Agnes Koschmider, Andreas Oberweis, Andreas Schoknecht, Meike Ullrich

BPM Academic Initiative: Fostering Academic Research in Business Process Management 21
Mathias Weske

The Free Models Initiative
Results of the First FMI Workshop, Vienna, March 18, 2014

Harald Störrle (Technical University of Denmark, Denmark)
Regina Hebig (Sorbonne Universites LIP6, Paris, France)
Alexander Knapp (University of Augsburg, Germany)

September 15, 2015

It is broadly accepted that collections of example data (¨Corpora¨) are an important driver for the progress of a scientific discipline. In the words of R. Dekker, "data-sets [...] are becoming more important themselves and can sometimes be seen as the primary intellectual output of the research" (cf. [1], p. 1). Model-based software development (MBSE) is no exception. For example, the quality of the used data strongly affects the reliability when new modeling techniques (e.g. approaches for model transformation, clone detection, or model-querying) are evaluated. Further, a comparison of competing approaches requires the application to comparable models. Therefore, a commonly accessible set of models will improve research. There are many situations where model corpora are helpful:

- Benchmarking: new approaches and algorithms ought to be validated against their predecessors to be able to accurately assess their contribution.
- Best Practices: model benchmarks and reference models may contribute to improving the state of the practice of modeling by making good (or bad) examples widely accessible.
- Validation: a body of examples that is generally accepted as being representative allows researchers to validate new models against them, as being equally valid in one aspect or another.

However, it seems that - despite the prominence of model repositories among researchers - it only seldom succeeds to publish and reuse real industrial models. Unlike other branches of science and engineering, software engineering (and in particular, MDSD) has not yet produced an accepted way of publishing models as data; there are no data journals and conferences. In fact, there are not many models freely available, and those that exist are hard to find, and not very rich in content. Of the few repositories in existence, most are relatively small and provide data without adequate meta-data or not in a machine-readable format. The purpose of this paper is to summarize the knowledge about existing model repositories, and distribute it to the community as an index to existing models.

The FMI 2014 workshop aimed at investigating reasons and exploring solutions for this situation. The discussions during the workshop raised additional questions and research challenges, that are summarized in the following.

The outcome of this workshop and the results of the work following-up the discussions is presented in this preface. Its contributions are:

– The Software Engineering Model Index (SEMI), a catalog of model repositories that we are building up currently. SEMI is supposed to serve as a common entry-point for researchers in need of models, and those that have access to models that they want to share. In Appendix A, we present the result of the collection at FMI'14 and the ensuing validation and preliminary assessments of the contents of the repository sightings.
– An overview over the use cases and challenges for model repositories. We hope to attract more contributions from the community to grow the index and encourage more researchers to release their models to the public domain.
– A collection of forms of Ersatz-Models that are worth to be considered as substitutes for real models when research is done.

1 Acknowledgements for Discussions and Pioneering

We are thankful to the participants of the FMI'14 workshop, for engaged and interesting discussions and their help to identify an initial list of repositories, namely Pit Pietsch, Andreas Schoknecht, Meike Ullrich, Christian Schneider, Bernhard Thalheim, Frank Wolff, Andreas Oberweis, Tom Thaler, and Mathias Weske.

Previously, several other initiatives did pioneering work and tried to stimulate the interest in publishing models and making them freely available, most notably the Open Models Initiative (www.openmodels.org), the Open Model Initiative (sic!, see www.openmodels.at), the BPM Academic Initiative (www.bpmail.org), and ReMODD (www.cs.colostate.edu/remodd/v1/). While none of them has truly changed the situation (more details on these are found below) these works are an inspiration for us to keep pushing for a changed culture of model publication and sharing.

2 Challenges

We seeded the discussion at the FMI'14 workshop with the following questions.

– Is there really a need for models, and if so, how large is the demand, what kinds of models are in demand?
– Acknowledging that existing model repositories are very restricted in scope, size, and quality of content, which improvements are the most pressing?
– One of the reasons there are no freely available models is the lack of an incentive for publishing them. Which incentives would be effective, appropriate and practical to overcome this impasse?
– What legal and technical obstacles impede model publication, and how can we overcome or circumvent them, i.e.: what licenses are useful? Can obfuscation be used to make more models available?
– How can models created in collaborations between academia and industry be added to the public domain? How can industrial partners be incentivized to release models? What experience and advice is available for such situations?

We expected that the workshop would only provide very partial answers to these questions, and, more likely, add new questions on top. From the resulting unordered collection, we distilled the following key challenges.

1. Archiving: The obvious technical challenge at the back-end is how to archive data with very high reliability, for very long time, yet readily accessible, and economically viable. This challenge has been addressed by others before, so we can probably rely on existing solutions and services such as ZENODO www.zenodo.org.
2. Access support: The front-end faces a less well explored challenge: how to search for models. Obviously, models need to be stored with metadata to be able to search for them in meaningful ways. But just which meta-data are sufficient to address the future (and thus unknown) needs of researchers with the effort of extracting meta-data from models? The right balance has yet to be found. Clearly, we should strive to extract as much meta-data from models automatically as is possible, but there are many formats and many information items that might be of interest.
3. Intellectual property: Models are intellectual property (IP), and many interesting models are developed in industrial co-operations which means that often industrial partners own the IP, or at least have a veto to publishing. On the one hand, this issue must be addressed by convincing industrial partners to accept co-operation agreements that allow the publication of models just like the publication of scientiffic articles is accepted today. One way of broadening the scope of publishable work is to offer obfuscation of models. How can usability of models for different research concerns be maintained, when model are obfuscated? Do model repositories need disclaimers to make researchers aware of threads to validity resulting from e.g. model collection or obfuscation? This is a topic that has not been the focus of much research.
4. Incentives: On the other hand, academic partners need to be incentivized to publish their models. If models were citable just like papers, and if publishing models were to receive recognition similar to publishing papers, we believe researchers would be motivated to contribute models when possible. Of course, the same recognition should be given on tenure approval committees and so on. So, in a nutshell, we are asking for no less than a cultural change in the community.

We believe that there is probably a mismatch between supply and demand of models: much more models are needed than are available. So, probably the biggest challenge is to find sufficiently many models to make the idea of sharing models practically useful for a large enough community of researchers.

3 Ersatz-models

If "real" models are indeed hard to procure and publish, the natural question to ask is how we can replace "real" models by Ersatz-models, begging the question what characterizes "real" models, and exactly which models can claim to be

"real". Clearly, the only sensible interpretation of "real" is "representative for a given purpose or question". In other words, as researchers, the only meaningful question is whether or not a given model sample is representative of the model population naturally occurring for the research question at hand.

So, for research about the typical size of models in industry, or which modeling elements are used how frequently in industry, industrial models are needed. Models arising from class assignments in academic teaching can not claim to be representative. Conversely, when asking for the most common modeling mistakes for novice modelers, using student's models are probably as good as those of novice modelers from industry.

3.1 Deriving Ersatz-models by obfuscation

There may be legal impediments to publishing industrial or academic models, e.g., the respective company wants to protect the intellectual property (IP) embodied in the model, or the copyright of a model is held by a student who disagrees with the publication. In these cases, we cannot publish the original model. However, if we can ensure that the IP is save, maybe we can convince the copyright holder to release the model anyway. One way of doing this is through model obfuscation.

An obfuscated model is one where the labels have been modified to hide their meaning. Obviously, some research objectives are not hindered by obfuscation, e.g., the size of models is very likely independent of the contents of element labels, while semantic model and/or element similarity might well depend on the actual label.

So, the research question has to be determined at the time of obfuscation which might severely restrict the applicability of the obfuscated models for other research. Likely, there are several obfuscation techniques of increasing power that create models of a decreasing degree of representativity, i.e., decreasing usefulness. At this point, it might be worth while pointing out that likely, there will be several model corpora, specializing for different kinds of research, and thus raising different requirements for obfuscation strength and/or method.

3.2 Re-modeling

If automatic obfuscation of a given original "real" model is not possible, a trusted person may manually obfuscate the model, or manually re-create a model (re-model) that is like the original with regards to the question at hand.[1] Clearly, such Ersatz-models are potentially perfectly representative of the original, but there is no direct way to establish for a third party such as a reviewer whether original and Ersatz-model actually are similar with respect to the question at hand. Re-created models may be even more realistic than "real" models, in some sense, since they are closer to the ideal type of a model. Questions may be raised, though, as to bias introduced by the modeler.

[1] The MOCA project at the University Siegen has re-created such models.

A special kind of re-created models are reference model, that are expressly created as ideal types of models for a given domain. If these reference models are created as a community effort and widely accepted as being representative, other form of remodeled-models might become in future, too.

3.3 Generated models

In some cases, Ersatz-models may be created automatically without being derived from any model at all. Consider, e.g., reverse-engineering models from source code: Large amounts of source code are publicly available in the form of open source projects. They can be used to easily and cheaply create large models automatically. Obviously, the model elements occurring in such models is not representative of, say, system analysis or requirements models. They might still be useful, however, to test the performance and scalability of a model querying approach. An example is the ID^2 model corpus which Bernhard Thalheim claims to posses, publication being prohibited by legal obstacles.

Alternatively, there are generators that create models of arbitrary size randomly. Again, they might be useful for performance analysis of algorithms, and they have the added advantage that, depending on the generator, they might actually preserve properties such as the model element type occurrence frequency.

4 Classification criteria

At this stage, no common framework for the classification of individual models, model families, and model repositories exist (we expect these to require different criteria in turn). So, the following is barely an initial collection that requires further discussion and, indeed, research.

- **Context:** First and foremost, it is necessary to include information about the model context, i.e., the usual model meta-data such as model purpose, or any auxiliary documents (e.g., reports).
- **Format:** Then, there is a need for technical data such as file format (XML, which XML version, XMI-version, which tool-specific variant of XMI).
- **Size:** Model size in terms of Kb, number of model elements, number of diagrams, number of individual models in a model family,
- **Language:** Modeling language (UML, BPMN, EPC, ...), version of modeling language, natural language used in labels and comments Model Family Previous/Subsequent versions of a given model, alternative models, modified models after QA exercises
- **Origin:** The original specification/set of requirements that was used to create the model, the source code used generate the model, the pictures/ documents used to harvest the model from
- **Provenience:** industrial/academic models, models from real life case studies or made-up case studies

– **Originator:** Who has recreated the models, what kind of profile and proficiency do the modeler(s) have, what tools and resource have been used, how long did it take?
– **Method:** Was any particular method used for (re-)creating the models, if so which?
– **Quality:** What quality are the models, what QA techniques have been applied with which result

Whether this list is exhaustive, and exactly which of these criteria is to be applied and/or needed is subject to future debate and research.

5 Roadmap

As a result of the FMI-Workshop, we define a roadmap and activities to further the initiative.

– Reach Out Raise the awareness and reach out to other/larger communities by publishing FMI at other venues, e.g., MODELS, ICSE, ASE. Develop and maintain a web site for the FMI as a common point of reference.
– Model Index Compile and maintain a list of known model repositories. Validate index entries, and certify their respective quality. Collect new sightings and develop universal model review criteria.
– Terminology Develop classification criteria to uniformly report individual models, families of models, and model repositories. Clarify the terminology (e.g., what is the difference between models, families of models, and model repositories).
– Model Observatory Develop the "Model Observatory" as an online system to support sharing models among researchers.
– Cultural Change Advocate a cultural change such that it becomes unacceptable to publish evaluations without publishing the models used in the evaluation.

As a first step, we are calling on everybody who knows about a model repository to share their knowledge and make it available to the scientific community by forwarding the information to us. We need the support from the community. Thus, we are reaching out to everybody to join this initiative, and come forward with their knowledge, model repositories (or individual models), and expertise: share your knowledge with the community, help with the online system, and input your expertise into the model assessment/review process!

6 Known Repositories

This appendix combines the results of the FMI'14 workshop and a subsequent validation and preliminary assessment by the authors. There are many references to existing model repositories, but frequently, there is little more evidence to their existence than hearsay. Privacy, broken links, and dead references in the

Repository	Size	Model Types	Access	Origin
ReMoDD http://www.cs.colstate.edu /remodd/v1/ thalb	60	Any models (PDF, XMI, ...)	Account	
Open Models Initiative http://openmodels.org/	70	Any models		
BPM Academic Initiative http://www.BPMAI.org	29285	Process Models (BPMN, EPCs, etc.) XML		
AtlanMod Metamodel Zoos http://www.emn.fr/z-info/atlanmod/index.php/Zoos	305	Meta Models (KM3, XMI, RDF, ...)	Free	
Versicherungs-Anwendungs-Architektur (VAA) http://www.gdv.de	90	Class & Use Case Models (PNG, Innovator)	Free	
Dutch Municipalities http://www.model-dsp.nl	>700	Process Models	Account	
eXperience http://www.experience-online.ch/cases/experience20.nsf/fallstudie.xsp	525	Case Studies (PDF)		
Reference Model Catalog http://rmk.iwi.uni-sb.de/catalog.php	2290	Model (Structured Abstracts)	Free	
404 **Insurance App. Architecture** http://www.ibm.com/solutions/sg/insurance/enterprise_aa/tech_details.htm	250	Product, Process & Information Models -	?	
404 **BIT Process Library** http://www.zurich.ibm.com/csc/bit/downloads.html	735	Process Models -	?	
404 **Suncorp-Metway Ltd** ?	>6000	Process Models -	Proprietary	
404 **SAP R/3 Reference Model** ?	>1000	Process Models ER/Org. Models -	Proprietary	

Fig. 1. Index of known model repositories

literature make it hard to verify the claims raised about them. We have collected the evidence and a preliminary validation of claims below.

- **Repository for Model Driven Development (ReMoDD)**
  `www.cs.colostate.edu/remodd/v1/` ReMoDD currently contains around 60 models in different modeling languages. The models are available for account holders, only. Models are stored in a large variety of formats, mostly PDF but also some in XMI.
- **Open Models Initiative (OMI)**
  `http://openmodels.org/` Like ReMoDD, OMI offers a platform allowing researchers to share models. There are currently around 70 models of different languages in the repository. The models are mostly available as pictures, some of them include other file formats like MDL. The access to the models is CC BY NA SA. In most cases no explicit hints indicating whether models stem from industry or not.
- **BPM Academic Initiative (BPM AI)**
  `http://www.BPMAI.org` The BPM AI is a platform for modeling and sharing models for teaching purposes. As of writing this, it claims to contain 29,285 process models in various machine-readable formats. Apparently, most of the models are created by students as part of their assignments, but some are motivated from industrial case studies, too.
- **AtlanMod Meta Model Zoos**
  `www.emn.fr/z-info/atlanmod/index.php/Zoos` This is a collection of around 305 meta models. Each of them is available in multiple formats (e.g. KM3, XMI, or RDF). The access to the models is free.
- **Versicherungsanwendungsarchitektur (VAA)**
  `www.gdvonline.de/vaa` The VAA is a standard from the association of German insurance industry. There are around 90 use case and class diagrams, most of them as diagrams in text documents, as PNG files, but also downloadable in INNOVATOR format. The access is free, the website itself is in German.
- **Dutch municipalities**
  `http://www.model-dsp.nl/` A large number of Dutch communes have created a common repository of communal administrative processes, which is said to contain 700-800 business process models. Access is restricted to registered members.
- **eXperience**
  `www.experience-online.ch/cases/experience20.nsf/fallstudie.xsp` eXperience is is a collection of 525 business modeling case studies, each of which is mainly a semi-structured text with a few embedded diagrams. Access is CC:BY NC. All case studies stem from industry, e.g., construction or electronics. The majority of the items in the collection are described in German.
- **IWi Reference Model Catalog (RMC)**, [5] The RMC contains structured meta-information about 2290 reference models, including the VAA and the IAA mentioned in this list. The RMC does not give access to the indexed models as such, but may help finding the models required for a particular

task. The meta-information is somewhat restricted, though, and seems to have not been updated since 2007.
- **Insurance Application Architecture (IAA)**, [3] The IAA is said to contain around 250 process models, but the link[2] reported in [3] is broken.
- **BIT Process Library**, [2] The BIT Process Library contains 735 process models according to [2], but the link[3] reported in [2] is broken. The collection has been cited several times.
- **Suncorp-Metway Ltd**, [4] The Suncorp process model repository for insurance processes contains over 6000, according to [4]. This is a purely proprietary corpus.
- **SAP R3 Process Reference Model** This model has been cited very many times, and although it is not in free circulation, there seem to be many copies.

We already identified more than a dozen further model collections and repository, of which many need to be checked for their status and content. Further, there are more repositories that we know of or have heard about, but could not verify. It seems that repositories sometimes get lost over time. Clearly, including such a reference in a research article is problematic, when claims are based on the availability of such model collections.

References

1. Ronald Dekker. The importance of having data-sets. In Proc. IATUL Conf. Purdue University, e-Pubs, 2006.
2. D. Fahland, C. Favre, B. Jobstmann, J. Koehler, N. Lohmann, H. Volzer, and K.Wolf. Instantaneous soundness checking of industrial business process models. In Proc. Intl. Conf. Business Process Management (BPM), volume 5701 of LNCS, pages 278-293. Springer, 2009.
3. Jochen M. Küster and N. N. Detecting and Resolving Process Model Differences in the Absence of a Change Log. In Proc. Intl. Conf. Business Process Modeling (BPM'08), number 0 in LNCS, pages 244-260. Springer, 2008.
4. M. La Rosa, Marlon Dumas, R. Uba, and Remco M. Dijkman. Business process model merging: An approach to business process consolidation. Technical report, QUT ePrints 38241, 2010.
5. Tom Thaler, Jrgen Walter, Peyman Ardalani, Peter Fettke, and Peter Loos. The Need for Process Model Corpora. In Proc. Intl. Ws. Free Models Initiative. DTU, 2014. Technical University of Denmark, DTU-TR-2014-15.

[2] `http://www.ibm.com/solutions/sg/insurance/enterprise_aa/tech_details.html`
[3] `http://www.zurich.ibm.com/csc/bit/downloads.html`

Experiences on the Quality and Availability of Test Models for Model Differencing Tools

Pit Pietsch*, Dennis Reuling*, Udo Kelter*, Jens Folmer†, Birgit Vogel-Heuser†
* Software Engineering Group
University of Siegen, Germany
{pietsch, dreuling, kelter}@informatik.uni-siegen.de
† Mechanical Engineering Department
Automation and Information Systems
Technische Universität München, Germany
{folmer, vogel-heuser}@ais.mw.tum.de

***Abstract*—In the last years Model Driven Engineering (MDE) became a central development paradigm in many application domains. Thus, many tools and methods were introduced to the community which subsequently have to be tested and evaluated. Unfortunately test models are not or only scarcely available for many of these domains. Even for domains where models are available, they often cannot be used for evaluation purposes; Either because they are represented in proprietary formats which cannot be processed by the tools or because they are of poor quality. In this position paper we discuss our experience from two different research projects. We'll share our experiences on the the availability and inadequacy of test models, as well as the experience we gained during the attempt to establish a benchmark set for differencing algorithms.**

I. Background

Model-Driven Development (MDD) became a central development paradigm in many application domains. In MDD models are the central artifacts of development and replace source code. The models themselves are collaboratively and concurrently developed by teams of modelers. Hence, the same problems as for the version management of source code arise for models, too. Because of this model versioning related tools were especially in the focus of research in the last couple of years.

Arguably the most important functionality in the context of model versioning is *model comparison*, i.e. the identification of common elements in two models as well as the edit operations which transformed the first model in the second. Model comparison is essential because it is a requirement for many advanced model versioning functionalities, e.g. difference visualization, 2-way or 3-way merge of models and model patching.

Our group is now working for more then 10 years in the context of model versioning and model comparison. We introduced the SiDiff Model Differencing Framework [6], [11], [3]. SiDiff is a generic, highly configurable tool set which can be adapted to any modeling domain to compute high quality difference. We have experiences both with partners in academia and industry[1]. Additionally we introduced the SiLift Difference Lifting Tool [4], [2], [5]. SiLift aims at lifting differences between models on to a more comprehensive level, which can be understand more easily.

In this position paper we address problems and obstacles we experienced in the context of the MOCA[2] and QuDiMo[3] research projects. In the context of MOCA we'll discuss our experiences about the availability and quality of test models needed, whereas in the context of QuDiMo we will present a failed attempt to establish a common benchmark set for model differencing algorithms.

[1]See http://www.sidiff.org/ for a list of partners.

II. Experiences

A. Availability of Test Models

The goal of the DFG project SPP1953 is to resolve problems which occur in long-living and continuous evolving software systems. Just like source code, models are developed collaboratively by teams of modelers and subject to continuous change. The specification and recognition of these changes is the key to understand and manage the evolution of model-based systems.

The MOCA project addresses this issue. In this context we developed SiLift, a tool which lifts low-level representations of changes to a higher abstraction level. This high-level representation is more comprehensible for the developers and in line with the editing behaviour they know from their modelling tools. To evaluate SiLift and proof its usefulness test models are needed. Such models should originate from a real world scenario and thus must be developed by domain experts.

While in this context real models in industrial contexts exist, these could not be made publicly available because they are regarded as corporate secrets. To this end an extensive and comprehensive case study was created by one of our project partners. This case study had to fulfill general requirements posed by the SPP1953, i.e. to be able to evaluate core elements

[2]MOCA is supported by the DFG (German Research Foundation) under the Priority Programme SPP1593: Design For Future – Managed Software Evolution

[3]The QuDiMo-Project is supported by the DFG (German Research Foundation) , 2010 - 2012, 2014 under grant 499/5-1

of software aspects by taking context, platform and software into account. To this end, the complexity of the case study had to be comparable to real industrial plants [12], so that SPP1593 projects are able to evaluate their different scientific methods based on realistic assumptions. Fundamentally, different versions of the software and their documentation are required by the projects. The projects have to be able to inspect and analyze the evolution of software variants.

The study is based on an existing pick and place unit (PPU), that is a manufacturing (discrete) process and used for teaching and research since 2001 at the Institute of Automation and Information Systems (Technische Universität München). Here, 15 different evolution scenarios (see [12]) were identified and the corresponding software models developed [1]. The software itself is based on programming languages commonly used in automation. Additionally, an UML derivate [14] was used as a programming language. The final case study is documented in a technical report [7]. This technical report includes information on the evolution scenarios as well as all relevant SysML models of the PPU.

In this project both sides, i.e. the domain experts and the tool developers, worked closely together during the creation of the case study. Requirements, expectations and problems from both sides were clearly communicated and a direct feedback loop existed from the get-go. In our opinion this is the most essential condition which ensured the quality of the created test models. Nonetheless, our partner had to invest a conceivable amount of time and effort in the creation of a case study while real models already existed.

While the mentioned case study is an example of a successful cooperation, we also made negative experiences in other projects. Models which were provided by (industry) partners often lacked the general quality [9], [10] to be used for evaluation of MDD tools. These models were often only small, oversimplified snapshots abstracting from the complexity of real models. Hence, it was not possible to evaluate our tools under realistic conditions, which inevitably lead to failed expectations on both sides. We also were involved in projects where partners provided models which were simply incorrect. The models contained syntactical or semantical errors and could not be processed by our tools. Furthermore, because the evolution and context of these test models was not documented, it is was not possible to improve their quality with justifiable effort.

B. A Benchmark Set for Model Differencing Algorithms

One of the objectives of the QuDiMo (German: Qualitätsoptimierte Differenzen für Modelle) research project is to empirically assess the quality and efficiency of model differencing algorithms. While several qualitative comparisons and assessments of the known approaches have been published, these assessments usually rely only on a functional analysis of the basic algorithms. There are virtually no comparisons which address non-functional properties. Available empirical evaluations have been conducted so far mostly by suppliers of the technologies, typically using a small set of use cases and data sets. They cannot be reproduced or repeated with competing approaches. Currently there are no standard benchmarks, challenges, test cases, or contests available which enable different approaches to be assessed on a common basis.

To address this shortcoming we initiated in 2012 issue of the Comparison and Versioning of Software Models (CVSM) workshop an initiative to establish a community driven benchmark set for differencing algorithms. We asked fellow researchers and practitioners to share their experiences and insights about common problems. The idea was to design an initial set of benchmarks, which were made publicly available and that tool developers can use them to evaluate and compare there algorithms. The general consensus during the workshop was that the missing objective evaluation of algorithms is indeed a problem. Many of the proposed algorithms have shortcomings which are not explicitly addressed by the evaluations presented in the accompanying papers. This is particularly problematic because many advanced model versioning functionalities, e.g. model merging and model patching, depend on high quality differences. A Call for Proposals were send out for the 2013 issue of the CVSM and various benchmarks and problematic examples were submitted, e.g. [8] and [13]. These submitted benchmarks were discussed by the community and a subset was selected, improved and published as the initial benchmark set. In this years issue of the CVSM we asked tool developers to submit there solutions to these benchmarks, but unfortunately none of the submissions to the workshop addressed the benchmark set.

One reason we identified so far why the community benchmark initiative failed is that some algorithms use proprietary formats to represent models and thus were not able to process the benchmark models (which were represented in the defacto standard EMF/Ecore) at all. Other algorithms were able to process the models, but they work only under very specific and strict assumptions on the development process and the modelling tools, e.g. that all model elements must have persistent identifiers, and therefore the algorithms could not use the test cases. Furthermore, the model differencing community is rather small, so that not many tools were addressed by the benchmark set to start with.

III. Expectations on the Workshop

We'd like to use the FMI workshop to share the insights we gained in our projects, with our tools and together with our partners from academia and industry. Particularly we would like to discuss the common problems we continuously encounter when we work with test models and why these models often are of poor quality. We also want to share which foundations and principles have to be met from our point of view so that partners from industry and academia can cooperate successfully. Finally we'd like to share our experience gained through the failed attempt to establish a community benchmark set for differencing algorithms.

REFERENCES

[1] Institute of Automation and Information Systems. The pick and place unit: Demonstrator for evolution in industrial plant automation. [online] http://www.ppu-demonstrator.org.

[2] T. Kehrer, U. Kelter, M. Ohrndorf, and T. Sollbach. Understanding model evolution through semantically lifting model differences with SiLift. In *Software Maintenance (ICSM), 2012 28th IEEE International Conference on*, pages 638–641, 2012.

[3] T. Kehrer, U. Kelter, P. Pietsch, and M. Schmidt. Adaptability of model comparison tools. In *Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering*, ASE 2012, pages 306–309, New York, NY, USA, 2012. ACM.

[4] T. Kehrer, U. Kelter, and G. Taentzer. A rule-based approach to the semantic lifting of model differences in the context of model versioning. In *Automated Software Engineering (ASE), 2011 26th IEEE/ACM International Conference on*, pages 163–172, 2011.

[5] T. Kehrer, U. Kelter, and G. Taentzer. Consistency-preserving edit scripts in model versioning. In *2013 IEEE/ACM 28th International Conference on Automated Software Engineering (ASE)*, pages 191–201, Nov. 2013.

[6] U. Kelter, J. Wehren, and J. Niere. A generic difference algorithm for uml models. In *Software Engineering 2005. Fachtagung des GI-Fachbereichs Softwaretechnik*, 2005.

[7] C. Legat, J. Folmer, and B. Vogel-Heuser. Evolution in industrial plant automation: A case study. In *Proceedings of the 39th Annual Conference of the IEEE Industrial Electronics Society*, IECON '13, Vienna, Austria, 2013.

[8] P. Pietsch, K. Müller, and B. Rumpe. Model matching challenge: Benchmarks for ecore and bpmn diagrams. *Softwaretechnik-Trends*, 33(2), 2013.

[9] P. Pietsch, H. S. Yazdi, and U. Kelter. Generating realistic test models for model processing tools. In *ASE*, pages 620–623, 2011.

[10] P. Pietsch, H. S. Yazdi, and U. Kelter. Controlled generation of models with defined properties. In *Software Engineering*, pages 95–106, 2012.

[11] C. Treude, S. Berlik, S. Wenzel, and U. Kelter. Difference computation of large models. In *ESEC-FSE '07: Proceedings of the the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*, pages 295–304, New York, NY, USA, 2007. ACM.

[12] B. Vogel-Heuser, J. Folmer, and C. Legat. Anforderungen an die softwareevolution in der automatisierung des maschinen- und anlagenbaus. *at - Automatisierungstechnik (accepted)*, 2014.

[13] M. Wimmer and P. Langer. A benchmark for model matching systems: The heterogeneous metamodel case. *Softwaretechnik-Trends*, 33(2), 2013.

[14] D. Witsch and B. Vogel-Heuser. Plc-statecharts: An approach to integrate uml-statecharts in open-loop control engineering-aspects on behavioral semantics and model-checking. In *18th World Congress of International Federation of Automation Control (IFAC)*, pages 7866–7872, Mailand, Italien, 2011.

The Need for Process Model Corpora

Tom Thaler, Jürgen Walter, Peyman Ardalani, Peter Fettke, Peter Loos

Institute for Information Systems (IWi) at the
German Research Center for Artificial Intelligence (DFKI) and
Saarland University
66123 Saarbrücken, Germany
firstname.lastname@iwi.dfki.de

Abstract. In spite of the current research activities developing methods and techniques for business process model analysis, a standardized and digital available process model corpus for evaluating these methods and techniques is still missing. Particularly with regard to a consistent appreciation of information systems such a corpus is of high importance, as it improves the development of standardized evaluations. The benefit of such corpora can also be observed in other fields of research like computational linguistics, biology, chemistry or medicine. Against that background the position paper at hand motivates the need for model corpora in general and process model corpora in particular. A short introduction on what the authors already did in terms of developing and establishing a model corpus enriches the paper. The current prototypical corpus version contains reference models, models from practice and models from controlled environments and comprises 16 model collections with 2290 process models.

1. Motivation

Nowadays companies use large model databases to manage their business process models, which serve as a knowledge base for the design of their information systems. Oftentimes, these databases contain several hundred or even thousands of models [1, 2], wherefore methods and techniques for complexity reduction, handling and analysis of these data are needed. This demand is explicitly addressed by the information systems research, e. g. in terms of process model similarity [1, 3], identification of structural analogies [4-6] or inductive reference modeling [7]. At the same time an access to real process models from practice is missing, which is often caused by legal aspects or privacy. Companies are afraid of losing their competitive advantage by the publication of their business processes. Indeed, there are several approaches focusing the conceptualization and the establishment of open access model repositories [8] (apromore.org, openmodels.org, openmodels.at, prozoom.ch) but concrete digital and processible models are very rare.

Already today some trends within the information system research in that direction can be observed, e. g. in terms of the interest of the Business Process Management Conference (BPM) in publishing the source code of software tools and implemented algorithms which are named in the proceedings. In that context, the possibility of replicating the published findings is of major interest. Nevertheless publishing the underlying data material is rarely focused. But particularly these data are essential for the replication and therefore of high importance for the research progress. The capabilities of corresponding corpora can be observed in different fields of research. E. g. the use of speech and text corpora in the fields of computational linguistics [9, 10] leaded to high benefits in speech processing, human computer interaction and automatic translation techniques. The use of genomic databases caused substantial progresses in biology, chemistry and medicine. Against that background, the authors already did a first step towards a process model corpus, which contains models in a standardized, digital and processible format.

2. Long-Term Research Objectives – A Vision

The authors' vision is developing a comprehensive model corpus which contains models in a standardized, digital and processible format. Thus, the following research objectives are focused: (1) Creating a consistent understanding of business application systems in different domains, (2) reusing the contained models in other contexts, (3) creating a homogeneous data basis for different application and analysis scenarios. The corpus should also be published for a free use in science. However, that highly depends on the license holder of the content which is contained in the corpus. Finally, the authors aim at publishing the corpus in terms of open models; similar to the open source idea, which was established in context of software development during the last years.

The initial point for that intention is the currently existing reference model catalogue [11] (rmk.iwi.uni-sb.de/). It contains 98 reference model entries with lexical data and meta-data like the number of containing single models. However, this catalogue does not contain digital processible models (in terms of the used

modeling language or a consistent exchange format) and there are also no entries on individual models from different domains.

Next to the mentioned practical aspects like the replication of research findings or the evaluation of methods, techniques and algorithms, theoretical questions can be addressed as well. Some examples are the creation of a consistent understanding of terms over different domains or the automatic identification of modeling rules and conventions while modeling. This may improve the further development of current modeling theories.

In order to present the range of applications and analysis, in the following the authors introduce some concrete scenarios. This overview is neither concluding nor comprehensive, but it should illustrate the benefit of model corpora for the information systems research.

- **Process Matching** describes the mapping of nodes of a process model to the nodes of another process model [12]. Corresponding approaches are used in context of model search, process model similarity, reusability of model fragments or inductive reference modeling. By using a process model corpus, the following question could be answered for instance: (1) To what extent automatic approaches are able to find matches which are manually determined. (2) Do elements or model fragments exist, which are available in several reference models?
- Analyzing **structural analogies** focusses the identification of similar or analogue structures within one or more models [4-6]. The following questions could be addressed: (1) Which structures can be observed frequently, which seldom? (2) Which structures can be observed in common? Do specific structure sequences exist? (3) Are there different structures in different domains? (4) Is it possible to define content independent process templates?
- A further scenario is the search of **process variants**, as there will likely be specific models available in different reference models, e. g. models related to acquisition and distribution or models in context of accounting. The automatic identification of such fragments or models would contribute the development of a comprehensive reference model over different domains. Independent from that, the corpus offers the possibility to reproduce the evolution of models, as model versions of different years can be analyzed.

3. The IWi Process Model Corpus

In order to give the presented vision a form, the authors use the method of vertical prototyping, whereby process models are focused with the Event-driven process chain as the central modelling language. The models and model collections added to the model corpus are derived from various sources, such as books, journals or conference proceedings as well as transcripts or audio recordings. Analogue and not processible sources were manually digitized using the software tool ARIS 7.2. In case of digitally available and processible sources, generally ProM 5.2 was used for transforming other model types to EPC. If that was not possible, the authors proceeded a manual transformation, whereby the transformation rules were formulated and documented. Furthermore, in some cases the models were adapted in order to provide a consistent and standardized corpus. These adaptions cover syntax and semantic corrections as well as the transformation of not supported EPC element types (in terms of the corpus), e. g. the SEQ connector, to alternative and supported constructs. Nevertheless, existing source files, e. g. petri-nets as PNML, were included to the corpus as well. An official publication of the developed corpus is currently not available, as legal aspects need to be clarified in future work.

Based on the origin and type, each model collection or each model within the developed model corpus could be allocated to exactly one of the following three categories:

- **Reference models:** Reference models generally consist of descriptive and prescriptive model elements [13]: In a descriptive sense, a reference model captures similarities of a category of companies. In a prescriptive sense, a reference model presents a proposal for the design of enterprises.
- **Individual models:** Individual models describe processes in specific organizations. These include business models as well as models in public administration.
- **Models from controlled modelling scenarios:** A situation based on a textual description will be modelled from different probands. This textual description helps the probands to have both a common understanding of the problem and a uniform terminology. Therefore, the resulting models are called controlled models.

Table 1 gives an overview on the models currently contained in the corpus. Also the category and the source of data as well as the nature of the source (analogue vs. digital) and its format (e. g. book, text, audio or file formats) are listed in the table. In addition, short descriptions, the national language (German and English) and the number of models in the respective model collection are presented. Spelling corrections and the introduction of the new German spelling rules were not considered as changes. On the other hand structural changes, such as the cutting or merging of certain models were considered as changes resulting from particular adaption rules. Most of the changes resulted from the correction of syntactic errors, such as the correction of missing events or functions, the correction of edges only having a start or end node, or the correction of events, which due to the

print version occurred twice. Subsequently, the corpus contains different model versions in order to address different analysis scenarios.

Table 1: Overview: developed process model corpus

C	Name | S | T | F	Remarks	L	#
R	ECO-Integral | [14] | a | book	1. Information systems for environmental management. Contains 38 EPCs and 11 function trees (as EPC).	de	49
		2. Contains EPCs only. Intermediate process interfaces are transformed into hierarchical functions. 3 EPCs are composed into one EPC for syntactical reasons.	de	36
R	Retail-H 1996 | [15] | a | book	1. Handelsinformationssysteme. Edition 1996. Contains 54 EPCs and 2 event hierarchies (as EPC).	de	56
		2. Correspondent to the first variant with transformed SEQ operators.	de	54
R	Retail-H 2004 | [16] | a | book	1. Handelsinformationssysteme. Edition 2004. Contains 58 EPCs and 2 event hierarchies (as EPC).	de	60
		2. Corresponds to first variant, but with transformed SEQ-Operator.	de	58
		3. Based on first variant with integrated event hierarchies and further structural adaptations.	de	58
		4. Based on second variant with integrated event hierarchies and further structural adaptations.	de	58
R	ITIL | Bought from Software AG | d | ARIS-DB	1. Reference model for the IT Service Management. Digitisation is based on [17-21] by the provider and contains 19 EPCs, with an example for explanation, and further 297 models of other types.	de en	19
R	SAP R/3 1998 | [22] | a | book	1. SAP R/3 reference model. Literal, syntactical and referencing errors corrected.	de	56
R	SAP R/3 | source unknown | d | EPML	1. SAP R/3 reference model with cryptic model names and without hierarchies.	en	604
		2. Added plain model names and hierarchies.	en	604
R	Y-CIM 2.1 | ARIS-Toolset | d | PDF	1. Reference model for industrial business processes. Contains the complete business model of the ARIS-Toolset 2.1a 1994 with syntactical corrections.	de	7
R	Y-CIM 1998 | [23] | a | book	1. Reference model for industrial business processes. Covers EPCs and function trees; inclusive exercise EPCs and descriptions.	de	55
		2. According to the first variant but without exercise EPCs and descriptions.	de	45
R	Y-CIM 1994 | [24] | a | book	1. Structural correspondent to the German Y-CIM 1998. Labels and model names come from [24].	en	55
		2. Adaptions according to the second German variant.	en	45
I	Custom B2B | s | a | Text	1. Processes describing software customizing and the production of special machinery.	de	46
I	Business registration | s | a / d | text and audio	1. Business registration processes of 8 German communes.	de	24
I	GK-Rewe | [25] | d | PDF	1. Basic course "*accounting"* at Chemnitz University.	de	34
		2. Syntactical errors corrected.	de	34
I	E-Payment | s | a | Text	1. Electronic payment process of governance.	de	38
I	PMC | [12] | d | PNML	1. Birth registration processes of 9 countries and University admission processes of 9 German Universities. Originally modeled as Petri-Nets. PNML files were transformed to EPCs with ProM.	en	18
		2. Some event nodes removed.	en	18
I	Vogelaar | [26] | d | PDF	1. Dutch governance processes. Originally modeled with YAWL. Transformed to EPCs using the transformation rules from the source document.	en	81
C	Exams | e | a | exam	1. Exams of a course at a German University between 2010 and 2012.	de	78
			Number of all models	**2290**

Legend: C: Category (R: reference model, I: individual model, C: controlled modeling); S: Source (s: self-created); T: Type of source (a: analogue, d: digital); F: format of source; L: national language (de: German, en: English); #: number of EPCs

4. Conclusion and Outlook

Altogether, the model corpus consists of 16 model collections with 2290 EPCs. In contrast to a (simple) single model the presented model corpus provides several models of different domains, sizes and national languages. Nevertheless, the developed model corpus is narrow in size in comparison to the domain at all. Thus, the corpus cannot be seen as representative. This can be drawn back to the availability of free accessible models. However, the model corpus can be used in a wide range of application scenarios. Thus, the authors have taken a first step towards the realization of the presented vision of an extensive model corpus. In contrast to existing approaches, the scientific need for concrete digitally processible models has been addressed, since in many cases a lack of a uniform data basis exists. The scope of the model corpus enables both the evaluation of existing algorithms, methods and techniques as well as their (further) development. Here, some possible application scenarios have been outlined briefly, which should be investigated in more detail in future work.

In addition to the application scenarios, the continuous development of the model corpus by adding further models (even by other researchers) is in the focus of further work. Moreover, the licensing issues that are associated with the provision of the model corpus have to be resolved, since this is the condition for a beneficial usability in the research community.

References

1. Dijkman, R., et al., Similarity of business process models: Metrics and evaluation. Information Systems, 2011. 36(2): p. 498-516.
2. Houy, C., et al. Business Process Management in the Large. in Business & Information Systems Engineering. 2011.
3. Mendling, J., Metrics for process models : empirical foundations of verification, error prediction, and guidelines for correctness. 2008: Springer.
4. Fettke, P. and P. Loos, Zur Identifikation von Strukturanalogien in Datenmodellen – Ein Verfahren und seine Anwendung am Beispiel des Y-CIM-Referenzmodells von Scheer. Wirtschaftsinformatik, 2005. 47(2): p. 89-100.
5. Ekanayake, C., et al., Approximate Clone Detection in Repositories of Business Process Models, in Business Process Management, A. Barros, A. Gal, and E. Kindler, Editors. 2012, Springer Berlin Heidelberg. p. 302-318.
6. Walter, J., P. Fettke, and P. Loos. Zur Identifikation von Strukturanalogien in Prozessmodellen. in Tagungsband der Multikonferenz Wirtschaftsinformatik (MKWI 2012). 2012. Braunschweig, Germany.
7. Ardalani, P., et al. Towards a Minimal Cost of Change Approach for Inductive Reference Model Development. in Proceedings of the 21st European Conference on Information Systems (ECIS 2013). 2013. Utrecht, Netherlands: AIS.
8. Koch, S., S. Strecker, and U. Frank, Conceptual Modelling as a New Entry in the Bazaar: The Open Model Approach, in Open Source Systems, IFIP 203, E. Damiani, et al., Editors. 2006, Springer: Berlin. p. 9-20.
9. Fellbaum, C., et al. WordNet: An Electronic Lexical Database. 1998 27.10.2010 [cited 2010 15.11.2010]; Available from: http://wordnet.princeton.edu/.
10. Kunze, C., Semantische Relationstypen in GermaNet, in Semantik im Lexikon, S. Langer and D. Schnorbusch, Editors. 2005, Narr. p. 161-178.
11. Fettke, P. and P. Loos, Der Referenzmodellkatalog als Instrument des Wissensmanagements - Methodik und Anwendung, in Wissensmanagement mit Referenzmodellen. Konzepte für die Anwendungssystem- und Organisationsgestaltung, J. Becker and R. Knackstedt, Editors. 2002, Springer: Berlin et al. p. 3-24.
12. Cayoglu, U., et al. The Process Model Matching Contest 2013. in 4th International Workshop on Process Model Collections: Management and Reuse (PMC-MR'13). 2013. Beijing.
13. Fettke, P. and J. vom Brocke, Referenzmodell, in Enzyklopädie der Wirtschaftsinformatik – Online-Lexikon. http://www.enzyklopaedie-der-wirtschaftsinformatik.de/, K. Kurbel, et al., Editors. 2008, Oldenbourg: München.
14. Krcmar, H., et al., eds. Informationssysteme für das Umweltmanagement - Das Referenzmodell ECO-Integral. 2000, Oldenbourg: München, Wien.
15. Becker, J. and R. Schütte, Handelsinformationssysteme. 1996, Landsberg/Lech: verlag moderne industrie.
16. Becker, J. and R. Schütte, Handelsinformationssysteme. Domänenorientierte Einführung in die Wirtschaftsinformatik. 2. ed. 2004, Frankfurt am Main: Redline Wirtschaft.
17. Office of Government Commerce, ITIL - Service Strategy. 2010, Norwich: TSO Information & Publishing Solutions.
18. Office of Government Commerce, ITIL - Service Design. 2010, Norwich: TSO Information & Publishing Solutions.
19. Office of Government Commerce, ITIL - Service Operation. 2010, Norwich: TSO Information & Publishing Solutions.
20. Office of Government Commerce, ITIL - Service Transition. 2010, Norwich: TSO Information & Publishing Solutions.
21. Office of Government Commerce, ITIL - Continual Service Improvement. 2010, Norwich: TSO Information & Publishing Solutions.
22. Keller, G. and T. Teufel, SAP R/3 prozeßorientiert anwenden – Iteratives Prozeß-Prototyping zur Bildung von Wertschöpfungsketten. 1998, Bonn et al.: Addison-Wesley.
23. Scheer, A.-W., Wirtschaftsinformatik - Referenzmodelle für industrielle Geschäftsprozesse [Studienausgabe]. 2. ed. 1998, Berlin et al.: Springer.
24. Scheer, A.-W., Business Process Engineering - Reference Models for Industrial Enterprises. 2. ed. 1994, Berlin et al.: Springer.
25. Kahlert, D. Grundkurs Rechnungswesen. 2010 [cited 2010 23.11.2010]; Available from: http://www.tu-chemnitz.de/wirtschaft/sapr3/gkrewe/epk/.
26. Vogelaar, J.J.C.L., et al., Comparing Business Processes to Determine the Feasibility of Configurable Models: A Case Study, in Business Process Management Workshops, LNBIP 100, F. Daniel, K. Barkaoui, and S. Dustdar, Editors. 2012, Springer: Berlin. p. 50-61.

Towards an Open Process Model Repository for Evaluations in Business Process Management Research

Agnes Koschmider, Andreas Oberweis, Andreas Schoknecht, and Meike Ullrich
Karlsruhe Institute of Technology (KIT),
D-76128 Karlsruhe, Germany
first_name.lastname@kit.edu

1 Necessity for a Business Process Model Repository

The acceptance of scientific work is favored by a solid empirical validation that investigates if research goals are met or whether the proposal outperforms other solutions regarding specific criteria. In the Business Process Management (BPM) community a great part of research centers around business process models. Current research topics include process model matching (or more generally process similarity), compliance checking of process models regarding regulatory rules or the discovery of process models from execution logs (Process Mining) to name just a few examples. Research in such areas would greatly benefit from an open and freely accessible model repository including a rich variety of exemplary business process models as used in practice. Such a set of models, which is accepted by the BPM community, would provide for *standardized*, *repeatable* and *comparable* experiments.

The following list (which does not aim to be complete) presents a few exemplary problems we have recognized and which could be mitigated by an open process model repository:

a) *Different characteristics of business process models and repositories*: According to our experiences there is a lack of standardized business process model sets or repositories. Generally it appears that evaluations in research publications related to business process models tend to use models of industry partners or generate models themselves without attaching these to the research results. This hampers the repeatability and comparability of evaluations in turn. When e.g. evaluating the performance of a query language for process model repositories a certain standardization of business process model characteristics (e.g., control flow structures, labeling style of elements) as well as repository characteristics (e.g., modelling language, amount of models) would foster the comparability with other query approaches. Also standardized query types would be beneficial for the comparison and repetition of evaluation results in this context.

b) *Repeatability of evaluations is complicated*: We realized for instance that various solutions in the process model matching area were proposed during the last few years, which provide evaluations with different business process model sets (e.g. company specific models which were not publicly available). In such a case the repeatability of the evaluation is greatly hampered. This could pose a problem for reviewers who want to reproduce the results or if independent researchers want to examine the evaluation in greater detail as e.g. in the context of an evaluation of an RDF store in the Semantic Web community [SGK+08].

c) *Difficulties when comparing different solutions for a problem*: The comparison of different solutions to a common problem in research related to business process models is difficult to achieve. The evaluations of these solutions will certainly use varying process model sets, different criteria or same criteria with varying parameters which aggravates comparisons. The Process Model Matching Contest[1] organized in conjunction with the BPM conference 2013

[1] http://processcollections.org/matching-contest

can be seen as a first step to a structured comparison of different solutions. The discussion of the contest results emphasized the problem of providing suitable models for the contest. We believe that other researchers do have the same problem of finding suitable models for their experiments. In our opinion the RDF Store Benchmark[2] is a very good example from which the BPM community could adopt some ideas.

In essence, we think that through an open and freely accessible process model repository combined with standardized evaluation frameworks *comparability* and *repeatability* of scientific evaluations in regard to business process models could be fostered. Besides, such a repository would facilitate students' works. Students writing a thesis could save a lot of effort retrieving suitable process models, hence they could focus on their research.

2 Existing Approaches

A few steps have been taken by the BPM community to provide freely accessible models and benchmark data sets. An already existing initiative is the Open Models Initiative.[3] Essentially, this initiative shares the same idea as the Free Models Initiative. It pursues an open source like approach to models from various areas with the ultimate goal of establishing a community that produces and shares models freely accessible [KGH07]. Through the website various modelling tools and community features are offered. Another example heading in the same direction is described in [FFO+12]. The authors propose the development of open reference enterprise models, which can be adopted by companies and advocates the establishment of an open modelling community.[4]

While we generally agree and support the requests for collaborative modelling and sharing of models, our idea differs in respect to the application domain of open models. We are not seeking models which can be used or adapted by practitioners but rather collections of models that can be utilized for evaluation purposes in BPM research. I.e. we are looking for artificial and real-life model datasets which are freely accessible and accepted by the BPM community (one example might be a dataset for the analysis and comparison of methods for calculating process model similarity). Consequently, such datasets would support the analysis and evaluation of research in the BPM area.

3 Challenges regarding the Realization

In this Section we want to propose a few aspects which should be addressed to realize a helpful open process model repository.

From an organizational perspective it seems important to launch a working group that promotes the idea and inspires and involves interested persons. This working group could provide centralized information related to repositories and model datasets, e.g. similar to the Petri Nets World website,[5] which provides an overview of Petri Net related sources. Such a website might attract people from both industry and academia who could provide process models. Right now there are multiple initiatives (e.g. Open Models Initiative, Apromore[6]) which maintain separate websites so it is difficult to obtain an overview of the different approaches.

Furthermore, it could be useful to additionally provide process logs to support the recently emerging Process Mining field. This research area is e.g. concerned with the discovery of models from event logs or the conformance of process executions with a prescriptive process model. Hence, research related to process models and research using process event logs is tightly connected and should be considered together in the initial phase of the Free Models Initiative. To this end, it would be beneficial to integrate the corresponding proposal presented in [RdMG+07].

Besides the organizational aspects, the technical details of the model repository have to be carefully designed as well in order to allow for an integration of various models with different characteristics in possibly heterogeneous formats. An appropriate storage solution combined with

[2] http://www.w3.org/wiki/RdfStoreBenchmarking
[3] http://www.openmodel.at
[4] http://openmodels.org/.
[5] http://www.informatik.uni-hamburg.de/TGI/PetriNets/index.html
[6] http://apromore.org/

a suitable classification scheme is required to facilitate the retrieval of specific models from the repository. Enhanced features could include e.g., sorting and tagging of models or a built-in model transformation to convert models to another process modeling notation. To this end, experiences from similar projects like e.g., Apromore should be taken into account. Regarding the quality of the existing process models in the repository, public comments coupled with version control features would support both discussions and changes. It has to be noted that in order to keep up real-world representativity, an enhancement of the existing models is actually not desired.

Yet another and purely technical solution – fundamentally different from the repository idea – is providing models for various purposes with the help of a model generator. Early examples stem from the Process Mining field with the development of process log generators (e.g. PLG [BS10] or SecSy [SA13]). Such generators should be able to produce models with various characteristics for different use cases. Possible parameters for such a generator could be number of activities, number of decisions, inclusion of certain workflow patterns, modelling language, etc. A possible use case might be the evaluation of process repository query language performance which might depend on the number of process models in a repository.

4 Conclusion

In this paper we elaborated on the hypothesis that rigorous research related to business process models would benefit from an open model repository. Through such a repository model datasets could be provided to the BPM community which would foster *standardized*, *repeatable* and *comparable* evaluations in research publications. To this end we suggest four ideas for future research directions: (i) instantiating a working group to provide for a centralized information source and contact point, (ii) integration of process execution logs into the repository, (iii) careful design of technical aspects of the repository and (iv) development of model generators.

Acknowledgement: This work has been developed with the support of DFG (German Research Foundation) under the project SemReuse OB 97/9-1.

References

[BS10] Andrea Burattin and Alessandro Sperduti. PLG: A Framework for the Generation of Business Process Models and Their Execution Logs. In *Business Process Management Workshops*, pages 214–219, 2010.

[FFO+12] Robert B. France, Ulrich Frank, Andreas Oberweis, Matti Rossi, and Stefan Strecker. Open Models as a Foundation of Future Enterprise Systems (Dagstuhl Seminar 12131). *Dagstuhl Reports*, 2(3):67–85, 2012.

[KGH07] Dimitris Karagiannis, Wilfried Grossmann, and Peter Höfferer. Open Model Initiative - A Feasibility Study. Technical report, University of Vienna, Departement of Knowledge and Business Engineering, 2007.

[RdMG+07] Anne Rozinat, Ana Karla Alves de Medeiros, Christian W. Günther, A. J. M. M. Weijters, and Wil M. P. van der Aalst. The Need for a Process Mining Evaluation Framework in Research and Practice. In *Business Process Management Workshops*, pages 84–89, 2007.

[SA13] Thomas Stocker and Rafael Accorsi. SecSy: Security-aware Synthesis of Process Event Logs. In *Workshop on Enterprise Modelling and Information Systems Architectures*, pages 71–84, 2013.

[SGK+08] Lefteris Sidirourgos, Romulo Goncalves, Martin Kersten, Niels Nes, and Stefan Manegold. Column-store Support for RDF Data Management: Not All Swans Are White. *Proceedings of the VLDB Endownment*, 1(2):1553–1563, 2008.

Talk

BPM Academic Initiative: Fostering Academic Research in Business Process Management

Mathias Weske

Hasso-Plattner-Institut fr Softwaresystemtechnik GmbH
Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam

This talk introduces the Business Process Management Academic Initiative, which is run by academics in the BPM field and which aims at stimulating education and research in this domain. To achieve its goals, the initiative provides several instruments: (i) A web-based modeling tool, which can be used free of charge by students and academic researchers. (ii) A rich set of teaching material in the BPM domain and (iii) a large set of process models to be used in empirical research. The talk discusses these aspects and also sketches the challenges and limitations of the initiative.